

# Geometry of faithfulness assumption in causal inference

Caroline Uhler<sup>1</sup> Garvesh Raskutti<sup>2</sup> Peter Bühlmann<sup>1</sup> Bin Yu<sup>2,3</sup>

<sup>1</sup>Seminar für Statistik  
ETH Zürich

<sup>2</sup>Department of Statistics, and

<sup>3</sup>Department of EECS  
UC Berkeley

## Abstract

Many algorithms for inferring causality rely heavily on the faithfulness assumption. The main justification for imposing this assumption is that the set of unfaithful distributions has Lebesgue measure zero, since it can be seen as a collection of hypersurfaces in a hypercube. However, due to sampling error the faithfulness condition alone is not sufficient for statistical estimation, and strong-faithfulness has been proposed and assumed to achieve uniform or high-dimensional consistency. In contrast to the plain faithfulness assumption, the set of distributions that is not strong-faithful has non-zero Lebesgue measure and in fact, can be surprisingly large as we show in this paper. We study the strong-faithfulness condition from a geometric and combinatorial point of view and give upper and lower bounds on the Lebesgue measure of strong-faithful distributions for various classes of directed acyclic graphs. Our results imply fundamental limitations for algorithms inferring causality based on partial correlations or on conditional independence testing in the Gaussian case.

## 1 Introduction

Determining causal structure among variables based on observational data is of great interest in many areas of science. While quantifying associations among variables is well-developed, inferring causal relations is a much more challenging task. A popular approach to make the causal inference problem more tractable is given by directed acyclic graph (DAG) models, which describe conditional dependence information and causal structure.

A DAG  $G = (V, E)$  consists of a set of vertices  $V$  and a set of directed edges  $E$  such that there is no directed cycle. We index  $V = \{1, 2, \dots, p\}$  and consider random variables  $\{X_i \mid i = 1, \dots, p\}$  associated to the nodes  $V$ . We denote a directed edge from vertex  $i$  to vertex  $j$  by  $(i, j)$  or  $i \rightarrow j$ . In this case  $i$  is called a *parent* of  $j$  and  $j$  is called a *child* of  $i$ . If there is a directed path  $i \rightarrow \dots \rightarrow j$ , then  $j$  is called a descendent of  $i$  and  $i$  an ancestor of  $j$ . The skeleton of a DAG  $G$  is the undirected graph obtained from  $G$  by substituting directed edges by undirected edges. Two nodes which are connected by an edge in the skeleton of  $G$  are called *adjacent*, and a triple of nodes  $(i, j, k)$  is an *unshielded triple* if  $i$  and  $j$  are adjacent

---

*Key words and phrases:* causal inference, PC-algorithm, (strong) faithfulness, conditional independence, directed acyclic graph, structural equation model, real algebraic hypersurface, Crofton's formula, algebraic statistics.

to  $k$  but  $i$  and  $j$  are not adjacent. An unshielded triple  $(i, j, k)$  is called a *v-structure* if  $i \rightarrow k$  and  $j \rightarrow k$ . In this case  $k$  is called a *collider*.

The problem of estimating a DAG from the observational distribution is ill-posed due to non-identifiability: in general, several DAGs encode the same conditional independence (CI) relations and therefore, the true underlying DAG cannot be identified from the observational distribution. However, assuming faithfulness (see Definition 1.1), the Markov equivalence class, i.e. the skeleton and the set of v-structures of a DAG, is identifiable [8, cf. Theorem 5.2.6], making it possible to infer some bounds on causal effects [7]. We focus here on the problem of estimating the Markov equivalence class of a DAG and argue that, even in the Gaussian case, severe complications arise for data of finite (or asymptotically increasing) sample size.

There has been a substantial amount of work on estimating the Markov equivalence class in the Gaussian case [3, 5, 10, 11]. Algorithms which are based on testing CI relations usually must require the faithfulness assumption [11, cf.]:

**Definition 1.1.** A distribution  $\mathbb{P}$  is *faithful* to a DAG  $G$  if no CI relations other than the ones entailed by the Markov property are present.

This means that if a distribution  $\mathbb{P}$  is faithful to a DAG  $G$ , all conditional (in-)dependences can be read-off from the DAG  $G$  using the so-called d-separation rule [11, cf.]. Two nodes  $i, j$  are *d-separated* given  $S$  if on every path between  $i$  and  $j$  there is either a non-collider which is in  $S$  or a collider including all its descendants which is not in  $S$ . For Gaussian models, the faithfulness assumption can be expressed in terms of the d-separation rule and conditional correlations as follows:

**Definition 1.2.** A multivariate Gaussian distribution  $\mathbb{P}$  is said to be faithful to a DAG  $G = (V, E)$  if for any  $i, j \in V$  and any  $S \subset V \setminus \{i, j\}$ :

$$j \text{ is d-separated from } i \mid S \iff \text{corr}(X_i, X_j \mid X_S) = 0.$$

The main justification for imposing the faithfulness assumption is that the set of unfaithful distributions to a graph  $G$  has measure zero. However, for data of finite sample size estimation error issues come into play. Robins et al. [10] showed that many causal discovery algorithms, and the PC-algorithm [11] in particular, are pointwise but not uniformly consistent under the faithfulness assumption. This is because it is possible to create a sequence of distributions that is faithful but arbitrarily close to an unfaithful distribution. As a result, Zhang and Spirtes [14] defined the strong-faithfulness assumption for the Gaussian case, which requires sufficiently large non-zero partial correlations:

**Definition 1.3.** Given  $\lambda \in (0, 1)$ , a multivariate Gaussian distribution  $\mathbb{P}$  is said to be  *$\lambda$ -strong-faithful* to a DAG  $G = (V, E)$  if for any  $i, j \in V$  and any  $S \subset V \setminus \{i, j\}$ :

$$j \text{ is d-separated from } i \mid S \iff |\text{corr}(X_i, X_j \mid X_S)| \leq \lambda.$$

The assumption of  $\lambda$ -strong-faithfulness is equivalent to requiring

$$\min\{|\text{corr}(X_i, X_j \mid X_S)|, j \text{ not d-separated from } i \mid S, \forall i, j, S\} > \lambda.$$

This motivates our next definition which is weaker than strong-faithfulness.

**Definition 1.4.** Given  $\lambda \in (0, 1)$ , a multivariate Gaussian distribution  $\mathbb{P}$  is said to be *restricted  $\lambda$ -strong-faithful* to a DAG  $G = (V, E)$  if both of the following hold:

- (i)  $\min\{|\text{corr}(X_i, X_j \mid X_S)|, (i, j) \in E, S \subset V \setminus \{i, j\} \text{ such that } |S| \leq \deg(G)\} > \lambda$ ,  
where here and in the sequel,  $\deg(G)$  denotes the maximal degree (i.e., sum of indegree and outdegree) of nodes in  $G$ ;
- (ii)  $\min\{|\text{corr}(X_i, X_j \mid X_S)|, (i, j, S) \in N_G\} > \lambda$ ,  
where  $N_G$  is the set of triples  $(i, j, S)$  such that  $i, j$  are not adjacent but there exists  $k \in V$  making  $(i, j, k)$  an unshielded triple, and  $i, j$  are not d-separated given  $S$ .

The first condition (i) is called *adjacency-faithfulness* in [15], the second condition (ii) is called *orientation-faithfulness*. If a multivariate Gaussian distribution  $\mathbb{P}$  satisfies adjacency-faithfulness with respect to a DAG  $G$ , we call the distribution  $\lambda$ -*adjacency-faithful* to  $G$ . Obviously, restricted  $\lambda$ -strong-faithfulness is a weaker assumption than  $\lambda$ -strong-faithfulness.

We now briefly discuss the relevance of these conditions and their use in previous work. Zhang and Spirtes [14] proved uniform consistency of the PC-algorithm under the strong-faithfulness assumption with  $\lambda \asymp 1/\sqrt{n}$ , for the low-dimensional case where the number of nodes  $p = |V|$  is fixed and sample size  $n \rightarrow \infty$ . In a high-dimensional and sparse setting, Kalisch and Bühlmann [5] require strong-faithfulness with  $\lambda_n \asymp \sqrt{\deg(G) \log(p)/n}$  (the assumption in [5] is slightly stronger, but can be relaxed as indicated here). Importantly, since  $\text{corr}(X_i, X_j \mid X_S)$  is required to be bounded away from 0 by  $\lambda$  for vertices that are not d-separated, the set of distributions that is not  $\lambda$ -strong-faithful no longer has measure 0.

It is easy to see for example from the proof in [5] that restricted  $\lambda$ -strong-faithfulness is a sufficient condition for consistency of the PC-algorithm in the high-dimensional scenario (with  $\lambda \asymp \sqrt{\deg(G) \log(p)/n}$ ) and that the condition is also sufficient and essentially necessary for consistency of the PC-algorithm. Furthermore, part (i) of the restricted strong-faithfulness condition is sufficient and essentially necessary for correctness of the conservative PC-algorithm [15], where correctness refers to the property that an oriented edge is correctly oriented but there might be some non-oriented edges which could be oriented (i.e., the conservative PC-algorithm may not be fully informative). The word “essentially” above means that we may consider too many possible separation sets  $S$  where  $|S| \leq \deg(G)$ , while the necessary collection of separating sets  $S$  which the (conservative) PC-algorithm has to consider might be a little bit smaller. Nevertheless, these differences are minor and we should think of part (i) of the restricted strong-faithfulness assumption as a necessary condition for consistency of the conservative PC-algorithm and both parts (i) and (ii) as a necessary condition for consistency of the PC-algorithm.

There are no known upper and lower bounds for the Lebesgue measure of  $\lambda$ -strong-unfaithful distributions or of restricted  $\lambda$ -strong-unfaithful distributions. Since these assumptions are so crucial to inferring structure in causal networks it is vital to understand if restricted and plain  $\lambda$ -strong-faithfulness are likely to be satisfied.

In this paper, we address the question of how restrictive the (restricted) strong-faithfulness assumption is using geometric and combinatorial arguments. In particular, we develop upper and lower bounds on the Lebesgue measure of Gaussian distributions that are not  $\lambda$ -strong-faithful for various graph structures. By noting that each CI relation can be written as

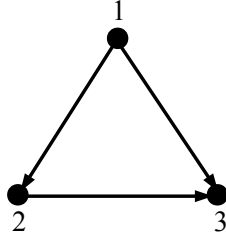


Figure 1: Motivating example: 3-node graph.

a polynomial equation and the unfaithful distributions correspond to a collection of real algebraic hypersurfaces, we exploit results from real algebraic geometry to bound the measure of the set of strong-unfaithful distributions. As we demonstrate in this paper, the strong-faithfulness assumption is restrictive for various reasons. Firstly, the number of hypersurfaces corresponding to unfaithful distributions may be quite large depending on the graph structure, and each hypersurface fills up space in the hypercube. Secondly, the hypersurfaces may be defined by polynomials of high degrees depending on the graph structure. The higher the degree, the greater the curvature and therefore the surface area of the corresponding hypersurface. Finally, to get the set of  $\lambda$ -strong-unfaithful distributions, these hypersurfaces get fattened up by a factor which depends on the size of  $\lambda$ .

Our results show that the set of distributions that do not satisfy strong-faithfulness can be surprisingly large even for small and sparse graphs (e.g. 10 nodes and an expected neighborhood (adjacency) size of 2) and small values of  $\lambda$  such as  $\lambda = 0.01$ . This implies fundamental limitations for algorithms based on partial correlations, with the PC-algorithm [11] as its most prominent example. As a consequence, other inference methods might be preferable which are not based on conditional independence testing (or partial correlation testing). The penalized maximum likelihood estimator [3] is such a method and consistency results without requiring strong-faithfulness have been given for the high-dimensional and sparse setting [13].

The remainder of this paper is organized as follows: Section 2 presents a simple example of a 3-node fully connected DAG, where we explicitly list the polynomial equations defining the hypersurfaces and plot the parameters corresponding to unfaithful distributions. In Section 3, we define the general model for a DAG on  $p$  nodes and give a precise description of the problem of bounding the measure of distributions that do not satisfy strong-faithfulness for general DAGs. In Section 4, we provide an algebraic description of the unfaithful distributions as a collection of hypersurfaces and give a combinatorial description of the defining polynomials in terms of paths along the graph. Section 5 provides a general upper bound on the measure of  $\lambda$ -strong-unfaithful distributions and lower bounds for various classes of DAGs, namely DAGs whose skeletons are trees, cycles or bipartite graphs  $K_{2,p-2}$ . Finally, in Section 6 we provide simulation results to validate our theoretical bounds.

## 2 Example: 3-node fully-connected DAG

In this section, we motivate the analysis in this paper using a simple example involving a 3-node fully-connected DAG. The graph is shown in Figure 1. We demonstrate that even in the 3-node case, the strong-faithfulness condition may be quite restrictive. We consider a Gaussian distribution which satisfies the directed Markov property with respect to the 3-node fully-connected DAG. An equivalent model formulation in terms of a Gaussian structural equation model is given as follows:

$$\begin{aligned} X_1 &= \epsilon_1 \\ X_2 &= a_{12}X_1 + \epsilon_2 \\ X_3 &= a_{13}X_1 + a_{23}X_2 + \epsilon_3, \end{aligned}$$

where  $(\epsilon_1, \epsilon_2, \epsilon_3) \sim \mathcal{N}(0, I)$ .<sup>1</sup> The parameters  $a_{12}, a_{13}$  and  $a_{23}$  reflect the causal structure of the graph. Whether the parameters are zero or non-zero determines the absence or presence of a directed edge.

It is well-known that through observing only covariance information it is not always possible to infer causal structure. In this example, the pairwise marginal and the conditional covariances are as follows:

$$\text{cov}(X_1, X_2) = a_{12} \tag{1}$$

$$\text{cov}(X_1, X_3) = a_{13} + a_{12}a_{23} \tag{2}$$

$$\text{cov}(X_2, X_3) = a_{12}^2a_{23} + a_{12}a_{13} + a_{23} \tag{3}$$

$$\text{cov}(X_1, X_2 \mid X_3) = a_{13}a_{23} - a_{12} \tag{4}$$

$$\text{cov}(X_1, X_3 \mid X_2) = -a_{13} \tag{5}$$

$$\text{cov}(X_2, X_3 \mid X_1) = -a_{23}. \tag{6}$$

If it were known a priori that the temporal ordering of the DAG is  $(X_1, X_2, X_3)$ , the problem of inferring the DAG-structure would reduce to a simple estimation problem. We would only need information about the (non-)zeroes of  $\text{cov}(X_1, X_2)$ ,  $\text{cov}(X_1, X_3 \mid X_2)$  and  $\text{cov}(X_2, X_3 \mid X_1)$ , that is, information whether the single edge weights  $a_{12}$ ,  $a_{13}$  and  $a_{23}$  are zero or not, which is a standard hypothesis testing problem. In particular, issues around (strong-) faithfulness would not arise. However, since the causal ordering of the DAG is unknown, algorithms based on conditional independence testing, which amount to testing partial correlations or conditional covariances, require that we check *all* partial correlations between two nodes given *any subset of remaining nodes*: a prominent example is the PC-algorithm [11]. For instance for the 3-node case, the PC-algorithm would infer that there is an edge between nodes 1 and 2 if and only if  $\text{cov}(X_1, X_2) \neq 0$  *and*  $\text{cov}(X_1, X_2 \mid X_3) \neq 0$ . The issue of faithfulness comes into play, because it is possible that all causal parameters  $a_{12}, a_{13}$  and  $a_{23}$  are nonzero while  $\text{cov}(X_1, X_2 \mid X_3) = 0$ , simply setting  $a_{12} = a_{13}a_{23}$  in (4).

Since in this example no CI relations are imposed by the Markov property, a distribution  $\mathbb{P}$  is unfaithful to  $G$  if any of the polynomials in (1)-(6) (corresponding to (conditional

---

<sup>1</sup>The assumption of  $\text{var}(\epsilon_j) \equiv 1$  is obviously restricting the class of Gaussian DAG models, but it does not affect issues with respect to strong-faithfulness.

# GEOMETRY OF FAITHFULNESS ASSUMPTION IN CAUSAL INFERENCE

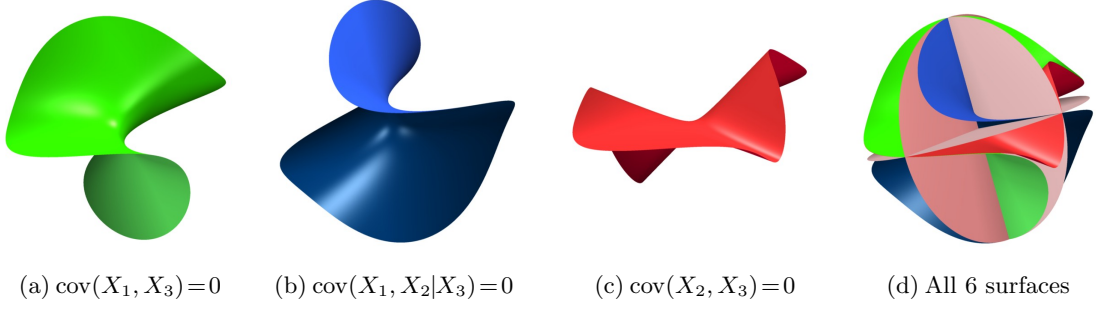


Figure 2: Parameter values corresponding to unfaithful distributions in the 3-node case.

covariances) are zero. Therefore, the set of unfaithful distributions for the 3-node example is the union of 6 real algebraic varieties, namely the three coordinate hyperplanes given by (1), (5) and (6), two real algebraic hypersurfaces of degree 2 given by (2) and (4), and one real algebraic hypersurface of degree 3 given by (3).

Assuming that the causal parameters lie in the cube  $(a_{12}, a_{13}, a_{23}) \in [-1, 1]^3$ , we use **surfex**, a software for visualizing algebraic surfaces, to generate a plot of the set of parameters leading to unfaithful distributions. Figure 2(a)-2(c) show the non-trivial hypersurfaces corresponding to  $\text{cov}(X_1, X_3) = 0$ ,  $\text{cov}(X_1, X_2 | X_3) = 0$  and  $\text{cov}(X_2, X_3) = 0$ . Figure 2(d) shows a plot of the union of all six hypersurfaces.

It is clear that the set of unfaithful distributions has measure zero. However, due to the curvature of the varieties and the fact that we are taking a union of 6 varieties, the chance of being "close" to an unfaithful distribution is quite large. As discussed earlier, being close to an unfaithful distribution is of great concern due to sampling error. Hence the set of distributions that does not satisfy  $\lambda$ -strong-faithfulness is of interest. As a direct consequence of Definition 1.3, this set of distributions corresponds to the set of parameters satisfying at least one of the following inequalities:

$$\begin{aligned} |\text{cov}(X_1, X_2)| &\leq \lambda \sqrt{\text{var}(X_1) \text{var}(X_2)}, \\ |\text{cov}(X_1, X_3)| &\leq \lambda \sqrt{\text{var}(X_1) \text{var}(X_3)}, \\ |\text{cov}(X_2, X_3)| &\leq \lambda \sqrt{\text{var}(X_2) \text{var}(X_3)}, \\ |\text{cov}(X_1, X_2 | X_3)| &\leq \lambda \sqrt{\text{var}(X_1 | X_3) \text{var}(X_2 | X_3)}, \\ |\text{cov}(X_1, X_3 | X_2)| &\leq \lambda \sqrt{\text{var}(X_1 | X_2) \text{var}(X_3 | X_2)}, \\ |\text{cov}(X_2, X_3 | X_1)| &\leq \lambda \sqrt{\text{var}(X_2 | X_1) \text{var}(X_3 | X_1)}. \end{aligned}$$

The set of parameters  $(a_{12}, a_{13}, a_{23})$  satisfying any of the above relations for  $\lambda \in (0, 1)$  has non-trivial volume. As we show in this paper, the volume of the distributions that are not  $\lambda$ -strong-faithful grows as the number of nodes and the graph density grow since both the number of varieties and the curvature of the varieties increase.

### 3 General problem setup

Consider a DAG  $G$ . Without loss of generality we assume that the vertices of  $G$  are *topologically ordered*, meaning that  $i < j$  for all  $(i, j) \in E$ . Each node  $i$  in the graph is associated with a random variable  $X_i$ . Given a DAG  $G$ , the random variables  $X_i$  are related to each other by the following structural equations:

$$X_j = \sum_{i < j} a_{ij} X_i + \epsilon_j, \quad j = 1, 2, \dots, p, \quad (7)$$

where  $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_p) \sim \mathcal{N}(0, I)$  (see footnote <sup>1</sup>) and  $a_{ij} \in [-1, +1]$  are the causal parameters with  $a_{ij} \neq 0$  if and only if  $(i, j) \in E$ . In matrix form, these equations can be expressed as

$$(I - A)^T X = \epsilon,$$

where  $X = (X_1, X_2, \dots, X_p)$  and  $A \in \mathbb{R}^{p \times p}$  is an upper triangular matrix with  $A_{ij} = a_{ij}$  for  $i < j$ . Since  $\epsilon \sim \mathcal{N}(0, I)$ ,

$$X \sim \mathcal{N}(0, [(I - A)(I - A)^T]^{-1}). \quad (8)$$

We will exploit the distributional form (8) for bounding the volume of the sets  $(a_{ij})_{(i,j) \in E} \in [-1, +1]^{|E|}$  that correspond to Gaussian distributions that are not (restricted)  $\lambda$ -strong-faithful.

Given  $(i, j) \in V \times V$  with  $i \neq j$  and  $S \subset V \setminus \{i, j\}$ , we define the set

$$\mathcal{P}_{i,j|S}^\lambda := \left\{ (a_{u,v}) \in [-1, +1]^{|E|} \mid |\text{cov}(X_i, X_j \mid X_S)| \leq \lambda \sqrt{\text{var}(X_i \mid X_S) \text{var}(X_j \mid X_S)} \right\}.$$

The set of parameters corresponding to distributions that are not  $\lambda$ -strong-faithful is

$$\mathcal{M}_{G,\lambda} := \bigcup_{\substack{i,j \in V, S \subset V \setminus \{i,j\}: \\ j \text{ not } d\text{-separated from } i \mid S}} \mathcal{P}_{i,j|S}^\lambda.$$

The set of parameters corresponding to distributions that are not restricted  $\lambda$ -strong-faithful is given by

$$\mathcal{N}_{G,\lambda}^{(1)} := \bigcup_{\substack{i,j \in V, S \subset V \setminus \{i,j\}: \\ (i,j,S) \notin N_G^{(1)}}} \mathcal{P}_{i,j|S}^\lambda,$$

where  $N_G^{(1)}$  denotes the set of triples  $(i, j, S)$ ,  $S \subset V \setminus \{i, j\}$  with  $|S| \leq \deg(G)$ , satisfying either  $(i, j) \in E$  or  $i, j$  are not  $d$ -separated given  $S$  and not adjacent but there exists  $k \in V$  making  $(i, j, k)$  an unshielded triple. The set of parameters corresponding to distributions that are not  $\lambda$ -adjacency-faithful (see part (i) of Definition 1.4) is given by

$$\mathcal{N}_{G,\lambda}^{(2)} := \bigcup_{\substack{i,j \in V, S \subset V \setminus \{i,j\}: \\ (i,j,S) \notin N_G^{(2)}}} \mathcal{P}_{i,j|S}^\lambda,$$

where  $N_G^{(2)}$  denotes the set of triples  $(i, j, S)$ ,  $S \subset V \setminus \{i, j\}$  with  $|S| \leq \deg(G)$ , satisfying  $(i, j) \in E$ .

Our goal is to provide upper and lower bounds on the volume of  $\mathcal{M}_{G,\lambda}$ ,  $\mathcal{N}_{G,\lambda}^{(1)}$  and  $\mathcal{N}_{G,\lambda}^{(2)}$  relative to the volume of  $[-1, 1]^{|E|}$ , that is, to provide upper and lower bounds for

$$\frac{\text{vol}(\mathcal{M}_{G,\lambda})}{2^{|E|}} \quad \text{and} \quad \frac{\text{vol}(\mathcal{N}_{G,\lambda}^{(1)})}{2^{|E|}} \quad \text{and} \quad \frac{\text{vol}(\mathcal{N}_{G,\lambda}^{(2)})}{2^{|E|}}.$$

This is the probability mass of  $\mathcal{M}_{G,\lambda}$ ,  $\mathcal{N}_{G,\lambda}^{(1)}$  and  $\mathcal{N}_{G,\lambda}^{(2)}$  if the parameters  $(a_{ij})_{(i,j) \in E}$  are distributed uniformly in  $[-1, +1]^{|E|}$ , which we will assume throughout the paper.

## 4 Algebraic description of unfaithful distributions

In this section, we first explain that the unfaithful distributions can always be described by polynomials in the causal parameters  $(a_{ij})_{(i,j) \in E}$  and therefore correspond to a collection of hypersurfaces in the hypercube  $[-1, +1]^{|E|}$ . We then give a combinatorial description of these defining polynomials in terms of paths in the underlying graph. The proofs can be found in Section 8.

**Proposition 4.1.** *Let  $i, j \in V$ ,  $S \subsetneq V \setminus \{i, j\}$  and  $Q = S \cup \{i, j\}$ . All CI relations in model (7) can be formulated as polynomial equations in the entries of the concentration matrix  $K = (I - A)(I - A)^T$ , namely:*

- (i)  $X_i \perp\!\!\!\perp X_j \iff (C(K))_{ij} = 0,$
- (ii)  $X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i, j\}} \iff K_{ij} = 0,$
- (iii)  $X_i \perp\!\!\!\perp X_j \mid X_S \iff \det(K_{Q^c Q^c})K_{ij} - K_{i Q^c} C(K_{Q^c Q^c}) K_{Q^c j} = 0,$

where  $C(B)$  denotes the cofactor matrix of  $B$ .<sup>2</sup>

We now give an interpretation of the polynomials defining the hypersurfaces corresponding to unfaithful distributions in directed Gaussian graphical models as paths in the skeleton of  $G$ . The concentration matrix  $K$  can be expanded as follows:

$$\begin{aligned} K &= (I - A)(I - A)^T \\ &= I - A - A^T + AA^T. \end{aligned}$$

This decomposition shows that the entry  $K_{ij}$ ,  $i \neq j$ , corresponds to the sum of all paths from  $i$  to  $j$  which lead over a collider  $k$  minus the direct path from  $i$  to  $j$  if  $j$  is a child of  $i$ , i.e.,

$$K_{ij} = \sum_{k: i \rightarrow k \leftarrow j} a_{ik}a_{jk} - a_{ij}. \quad (9)$$

Note that  $a_{ij}$  is zero in the case that  $j$  is not a child of  $i$ .

---

<sup>2</sup>The  $(i, j)$ th cofactor is defined as  $C(K)_{ij} = (-1)^{i+j} M_{ij}$  where  $M_{ij}$  is the  $(i, j)$ th minor of  $K$ , i.e.,  $M_{ij} = \det(A(-i, -j))$ , where  $A(-i, -j)$  is the submatrix of  $A$  obtained by removing the  $i$ th row and  $j$ th column of  $A$ .



For the covariance matrix  $\Sigma = K^{-1}$  the equivalent result describing the path interpretation is given in [12, Equation (1)], namely

$$\Sigma = \sum_{k=0}^{2p-2} \sum_{\substack{r+s=k \\ r,s \leq p-1}} (A^T)^r A^s. \quad (10)$$

We give a proof using Neumann power series in Section 8.

Equation (10) shows that the  $(i, j)$ -th entry of  $\Sigma$  corresponds to all paths from  $i$  to  $j$ , which first go backwards until they reach some vertex  $k$  and then forwards to  $j$ . Such paths are called *treks* in [12]. In other words,  $\Sigma_{ij}$  corresponds to all collider-free paths from  $i$  to  $j$ .

We now understand the covariance between two variables  $X_i$  and  $X_j$  and the conditional covariance when conditioning on all remaining variables in terms of paths from  $i$  to  $j$ . In the following, we will extend these results to conditional covariances between  $X_i$  and  $X_j$  when conditioning on a subset  $S \subsetneq V \setminus \{i, j\}$ . This means that we need to find a path description of

$$P_{ij|S} := \det(K_{Q^c Q^c}) K_{ij} - K_{i Q^c} C(K_{Q^c Q^c}) K_{Q^c j} \quad (11)$$

(see Proposition 4.1 (iii)) and therefore of the determinant and the cofactors of  $K_{Q^c Q^c}$ .

Ponstein [9] gave a beautiful path description of  $\det(\lambda I - M)$  and the cofactors of  $\lambda I - M$ , where  $M$  denotes a variable adjacency matrix of a not necessarily acyclic directed graph. By replacing  $M$  by  $A + A^T - AA^T$ , that is by symmetrizing the graph and reweighting the directed edges, we can apply Ponstein's theorem.

**Ponstein's theorem.** *Let  $i, j \in V$ ,  $S \subsetneq V \setminus \{i, j\}$  and  $Q = S \cup \{i, j\}$  and let  $\hat{G}$  denote the weighted directed graph corresponding to the adjacency matrix  $A + A^T - AA^T$  and  $\hat{G}_{Q^c}$  the subgraph resulting from restricting  $\hat{G}$  to the vertices in  $Q^c$ . Then:*

$$(i) \det(K_{Q^c Q^c}) = 1 + \sum_{k=1}^{|Q^c|} \sum_{m_1 + \dots + m_s = k} (-1)^s \mu(c_{m_1}) \cdots \mu(c_{m_s}),$$

$$(ii) (C(K_{Q^c Q^c}))_{ij} = \sum_{k=2}^{|Q^c|} \sum_{m_0 + \dots + m_s = k-1} (-1)^s \mu(d_{m_0}) \mu(c_{m_1}) \cdots \mu(c_{m_s}), \text{ for } i \neq j,$$

where  $\mu(d_{m_0})$  denotes the product of the edge weights along a self-avoiding path from  $i$  to  $j$  in  $\hat{G}_{Q^c}$  of length  $m_0$ ,  $\mu(c_{m_1}), \dots, \mu(c_{m_s})$  denote the product of the edge weights along self-avoiding cycles in  $\hat{G}_{Q^c}$  of lengths  $m_1, \dots, m_s$ , respectively, and  $d_{m_0}, c_{m_1}, \dots, c_{m_s}$  are disjoint paths.

Putting together the various pieces in (11), namely Equation (9) for describing  $K_{QQ}$ ,  $K_{QQ^c}$  and  $K_{Q^c Q}$ , and Ponstein's Theorem for  $\det(K_{Q^c Q^c})$  and  $C(K_{Q^c Q^c})$ , we get a path interpretation of all partial correlations.

**Example 4.2.** For the special case where the underlying DAG is fully connected and we condition on all but one variable, i.e.,  $S = V \setminus \{i, j, s\}$ , the representation of the conditional

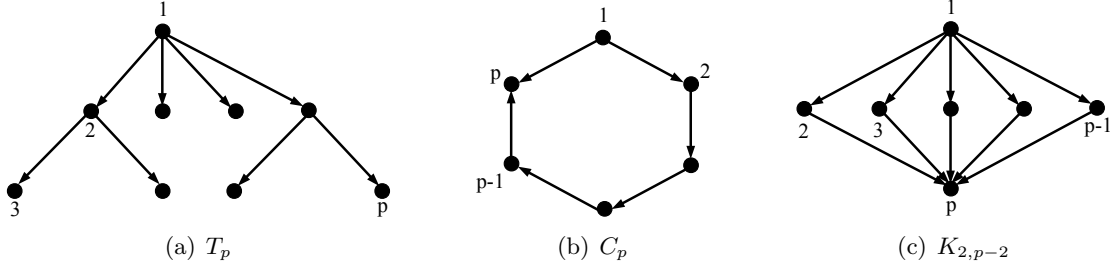


Figure 3: Directed tree, cycle and bipartite graph.

correlation between  $X_i$  and  $X_j$  when conditioning on  $X_S$  in terms of paths in  $G$  is given by

$$\left(1 + \sum_{k: s \rightarrow k} a_{sk}^2\right) \left(\sum_{k: i \rightarrow k \leftarrow j} a_{ik}a_{jk} - a_{ij}\right) - \left(\sum_{t: i \rightarrow t \leftarrow s} a_{it}a_{st} - a_{is}\right) \left(\sum_{t: j \rightarrow t \leftarrow s} a_{jt}a_{st} - a_{js}\right).$$

In the following, we apply Equation (9), Equation (10) and Ponstein's Theorem to describe the structure of the polynomials corresponding to unfaithful distributions for various classes of DAGs, namely DAGs whose skeletons are trees, cycles and bipartite graphs. We denote by  $T_p$  a directed connected rooted tree on  $p$  nodes, where all edges are directed away from the root as shown in Figure 3(a). Let  $C_p$  denote a DAG whose skeleton is a cycle, and  $K_{2,p-2}$  a DAG whose skeleton is a bipartite graph, where the edges are directed as shown in Figure 3(b) and Figure 3(c).

We denote by  $SOS(a)$  a *sum of squares* polynomial in the variables  $(a_{ij})_{(i,j) \in E}$ , meaning

$$SOS(a) = \sum_k f_k^2(a),$$

where each  $f_k(a)$  is a polynomial in  $(a_{ij})_{(i,j) \in E}$ . The polynomials corresponding to unfaithful distributions for the graphs described in Figure 3 are given in the following result.

**Corollary 4.3.** *Let  $i, j \in V$  and  $S \subset V \setminus \{i, j\}$  such that  $i, j$  are not  $d$ -separated given  $S$ . Then the polynomials  $P_{ij|S}$  defined in (11) corresponding to the CI relation  $X_i \perp\!\!\!\perp X_j \mid X_S$  in model (7) are of the following form:*

(a) for  $G = T_p$ :

$$a_{i \rightarrow j} \cdot (1 + SOS(a)),$$

where  $a_{i \rightarrow j}$  is a monomial and denotes the value of the unique path from  $i$  to  $j$ ;

(b) for  $G = C_p$ :

$$\begin{aligned} & a_{i \rightarrow j} \cdot (1 + SOS(a)) && \text{if } p \notin S, \\ & f(\bar{a})a_{i,i+1} - g(\bar{a})a_{j,j+1} && \text{if } S = \{p\}, \end{aligned}$$

where  $a_{i \rightarrow j}$  denotes the value of a path from  $i$  to  $j$  and  $f(\bar{a}), g(\bar{a})$  are polynomials in the variables  $\bar{a} = \{a_{st} \mid (s, t) \notin \{(i, i+1), (j, j+1)\}\}$ ;

(c) for  $G = K_{2,p-2}$ :

$$\begin{aligned} a_{i \rightarrow j} \cdot (1 + \text{SOS}(a)) & \quad \text{if } p \notin S, \\ f(\bar{a})a_{1,j} - g(\bar{a})a_{j,p} & \quad \text{if } i = 1 \text{ and } p \in S. \end{aligned}$$

## 5 Bounds on the volume of unfaithful distributions

Based on the path interpretation of the partial covariances explained in the previous section, we derive upper and lower bounds on the volume of the parameters that lead to  $\lambda$ -strong-unfaithful distributions. We also provide bounds on the proportion of restricted  $\lambda$ -strong-unfaithful distributions. These are distributions which do not satisfy the necessary conditions for uniform or high-dimensional consistency of the PC-algorithm. Our first result makes use of Crofton's formula for real algebraic hypersurfaces and the Lojasiewicz inequality to provide a general upper bound on the measure of strong-unfaithful distributions.

Crofton's formula gives an upper bound on the surface area of a real algebraic hypersurface defined by a degree  $d$  polynomial, namely:

**Crofton's formula.** *The volume of a degree  $d$  real algebraic hypersurface in the unit  $m$ -ball is bounded above by  $C(m)d$ , where  $C(m)$  satisfies*

$$\binom{m+d}{d} - 1 \leq C(m) d^m.$$

For more details on Crofton's formula for real algebraic hypersurfaces see for example [2] or [4, pages 45-46].

The Lojasiewicz inequality gives an upper bound for the distance of a point to the nearest zero of a given real analytic function. This is used as an upper bound for the thickness of the fattened hypersurface.

**Lojasiewicz inequality.** *Let  $f : \mathbb{R}^p \rightarrow \mathbb{R}$  be a real-analytic function and  $K \subset \mathbb{R}^p$  compact. Let  $V_f \subset \mathbb{R}^p$  denote the real zero locus of  $f$ , which is assumed to be non-empty. Then there exist positive constants  $c, k$  such that for all  $x \in K$ :*

$$\text{dist}(x, V_f) \leq c|f(x)|^k.$$

**Theorem 5.1** (General upper bound). *Let  $G = (V, E)$  be a DAG on  $p$  nodes. Then*

$$\begin{aligned} \frac{\text{vol}(\mathcal{N}_{G,\lambda}^{(2)})}{2^{|E|}} & \leq \frac{\text{vol}(\mathcal{N}_{G,\lambda}^{(1)})}{2^{|E|}} \leq \frac{\text{vol}(\mathcal{M}_{G,\lambda})}{2^{|E|}} \\ & \leq \frac{C(|E|)c\kappa^k\lambda^k}{2^{\frac{|E|}{2}}} \sum_{i,j \in V} \sum_{S \subset V \setminus \{i,j\}} \deg(\text{cov}(X_i, X_j \mid X_S)), \end{aligned}$$

# GEOMETRY OF FAITHFULNESS ASSUMPTION IN CAUSAL INFERENCE

where  $C(|E|)$  is a positive constant coming from Crofton's formula,  $c, k$  are positive constants, depending on the polynomials characterizing exact unfaithfulness (for an exact definition, see the proof), and  $\kappa$  denotes the maximal partial variance over all possible parameter values  $(a_{st}) \in [-1, 1]^{|E|}$ , i.e.,

$$\kappa = \max_{i,j \in V, S \subset V \setminus \{i,j\}} \max_{(a_{st}) \in [-1, 1]^{|E|}} \text{var}(X_i \mid X_S).$$

Theorem 5.1 shows that the volume of (restricted)  $\lambda$ -strong-unfaithful distributions may be large for two reasons. Firstly, the number of polynomials grows quickly as the size and density of the graph increases, and secondly the degree of the polynomials grows as the number of nodes and density of the graph increases. The higher the degree, the greater the curvature of the variety and hence the larger the volume that is filled according to Crofton's formula. Unfortunately, the upper bound cannot be computed explicitly, since we do not have bounds on the constants in the Lojasiewicz inequality.

*Proof.* It is clear that

$$\text{vol}(\mathcal{N}_{G,\lambda}^{(2)}) \leq \text{vol}(\mathcal{N}_{G,\lambda}^{(1)}) \leq \text{vol}(\mathcal{M}_{G,\lambda}).$$

Using the standard union bound we get that

$$\text{vol}(\mathcal{M}_{G,\lambda}) \leq \sum_{\substack{i,j \in V, S \subset V \setminus \{i,j\}: \\ j \text{ not } d\text{-separated from } i \mid S}} \text{vol}(\mathcal{P}_{ij|S}^\lambda).$$

Let  $V_{ij|S}$  denote the real algebraic hypersurface defined by  $\text{cov}(X_i, X_j \mid X_S)$ , i.e., the set of all parameter values  $(a_{st}) \in [-1, +1]^{|E|}$  which vanish on  $\text{cov}(X_i, X_j \mid X_S)$ . Hence,

$$\begin{aligned} \text{vol}(\mathcal{P}_{ij|S}^\lambda) &\leq \text{vol}(\{(a_{st}) \in [-1, +1]^{|E|} \mid |\text{cov}(X_i, X_j \mid X_S)| \leq \lambda \kappa\}) \\ &\leq \text{vol}(\{(a_{st}) \in [-1, +1]^{|E|} \mid \text{dist}((a_{st}), V_{ij|S}) \leq c_{ij|S} \lambda^{k_{ij|S}} \kappa^{k_{ij|S}}\}), \end{aligned}$$

where  $c_{ij|S}, k_{ij|S}$  are positive constants and the second inequality follows from the Lojasiewicz inequality.

We apply Crofton's formula on an  $|E|$ -dimensional ball of radius  $\sqrt{2}$  to get an upper bound on the surface area of a real algebraic hypersurface in the hypercube  $[-1, 1]^{|E|}$ :

$$\text{vol}(\mathcal{P}_{ij|S}^\lambda) \leq c_{ij|S} \lambda^{k_{ij|S}} \kappa^{k_{ij|S}} 2^{\frac{|E|}{2}} C(|E|) \deg(\text{cov}(X_i, X_j \mid X_S)).$$

The claim follows by setting

$$c = \max_{i,j \in V, S \subset V \setminus \{i,j\}} c_{ij|S} \quad \text{and} \quad k = \max_{i,j \in V, S \subset V \setminus \{i,j\}} k_{ij|S}.$$

□

The PC-algorithm in practice only requires  $\lambda$ -strong-faithfulness for all subsets  $S \subset V \setminus \{i, j\}$  for which  $|S|$  is at most the maximal degree of the graph. This could lead to a tighter upper bound, since we have fewer summands. We will analyze in Section 6 how helpful this is in practice.

Since the main goal of this paper is to show how restrictive the (restricted) strong-faithfulness assumption is, lower bounds on the proportion of (restricted)  $\lambda$ -strong-unfaithful distributions are necessary. However, non-trivial lower bounds for general graphs cannot be found using tools from real algebraic geometry, since in the worst case the surface area of a real algebraic hypersurface is zero. This is the case when the polynomial defining the hypersurface has no real roots. In that case the corresponding real algebraic hypersurface is empty. As a consequence, we need to analyze different classes of graphs separately, understand the defining polynomials, and find lower bounds for these classes of graphs. In Section 4, we discussed the structure of the defining polynomials for DAGs whose skeleton are trees, cycles or bipartite graphs, respectively. In the following, we use these results to find lower bounds on the proportion of (restricted)  $\lambda$ -strong-unfaithful distributions for these classes of graphs.

**Theorem 5.2** (Lower bound for trees). *Let  $T_p$  be a connected directed tree on  $p$  nodes with edge set  $E$  as shown in Figure 3(a). Then*

$$\begin{aligned} (i) \quad & \frac{\text{vol}(\mathcal{M}_{T_p, \lambda})}{2^{|E|}} \geq 1 - (1 - \lambda)^{p-1}, \\ (ii) \quad & \frac{\text{vol}(\mathcal{N}_{T_p, \lambda}^{(1)})}{2^{|E|}} \geq 1 - (1 - \lambda)^{p-1}. \\ (iii) \quad & \frac{\text{vol}(\mathcal{N}_{T_p, \lambda}^{(2)})}{2^{|E|}} \geq 1 - (1 - \lambda)^{p-1}. \end{aligned}$$

Theorem 5.2 shows that the measure of restricted and ordinary  $\lambda$ -strong-unfaithful distributions converges to 1 exponentially in the number  $p$  of nodes for fixed  $\lambda \in (0, 1)$ . Hence, even for trees the strong-faithfulness assumption is restrictive and the use of the PC-algorithm problematic when the number of nodes is large.

*Proof.* (i) For a given pair of nodes  $i, j \in V$ ,  $i \neq j$ , and subset  $S \subset V \setminus \{i, j\}$  we want to lower bound the volume of parameters  $(a_{st}) \in [-1, 1]^{|E|}$  (in this example  $|E| = p - 1$ ) for which

$$|\text{cov}(X_i, X_j \mid X_S)| \leq \lambda \sqrt{\text{var}(X_i \mid X_S) \text{var}(X_j \mid X_S)}$$

or equivalently

$$|P_{ij|S}| \leq \lambda \sqrt{P_{ii|S} P_{jj|S}}.$$

From Corollary 4.3 we know that the defining polynomials  $P_{ij|S}$  for  $T_p$  are of the form

$$a_{i \rightarrow j} \cdot (1 + \text{SOS}(a)).$$

Similarly as in Corollary 4.3 one can prove that the polynomials  $P_{ii|S}$  are of the form  $1 + \text{SOS}(a)$  and can therefore be lower bounded by 1.

## GEOMETRY OF FAITHFULNESS ASSUMPTION IN CAUSAL INFERENCE

So the hypersurfaces representing the unfaithful distributions are the coordinate planes corresponding to the  $p - 1$  edges in the tree  $T_p$ . A distribution is strong-unfaithful if it is near to any one of the hypersurfaces (worst case). Since there is a defining polynomial  $P_{ij|S}$  without the factor consisting of the sum of squares, the  $\lambda$ -strong-unfaithful distributions correspond to the parameter values  $(a_{st}) \in [-1, 1]^{p-1}$  satisfying

$$|a_{i \rightarrow j}| \leq \lambda$$

for at least one pair of  $i, j \in V$ . Since we are seeking a lower bound, we set all parameter values to 1 except for one. As a result, a lower bound on the proportion of  $\lambda$ -strong-unfaithful distributions is given by the union of all parameter values  $(a_{st}) \in [-1, 1]^{p-1}$  such that

$$|a_{st}| \leq \lambda.$$

We get a lower bound on the volume by an inclusion-exclusion argument. We first sum over the volume of all by  $2\lambda$  thickened coordinate hyperplanes, subtract all pairwise intersections, add all three-wise intersections, and so on. This results in the following lower bound:

$$\begin{aligned} \frac{\text{vol}(\mathcal{M}_{T_p, \lambda})}{2^{|E|}} &\geq (p-1) \frac{2\lambda 2^{p-2}}{2^{p-1}} - \binom{p-1}{2} \frac{(2\lambda)^2 2^{p-3}}{2^{p-1}} + \dots \\ &= \sum_{k=1}^{p-1} (-1)^{k+1} \binom{p-1}{k} \lambda^k \\ &= 1 - \sum_{k=0}^{p-1} \binom{p-1}{k} (-\lambda)^k \\ &= 1 - (1 - \lambda)^{p-1}. \end{aligned}$$

The proof of (ii) and (iii) is similar. The monomials  $a_{i \rightarrow j}$  reduce to single parameters  $a_{ij}$ , since the necessary conditions only involve  $(i, j) \in E$ .  $\square$

This theorem is in line with the results in [1], where they show that for trees checking if a Gaussian distribution satisfies all conditional independence relations imposed by the Markov property only requires testing if the causal parameters corresponding to the edges in the tree are non-zero.

Note that the behavior stated in Theorem 5.2 is qualitatively the same as for a linear model  $Y = X\beta + \epsilon$  with active set  $S = \{j \mid \beta_j \neq 0\}$ . To get consistent estimation of  $S$ , a “beta-min” condition is required, namely that for some suitable  $\lambda$ ,

$$\min_{j \in S} |\beta_j| > \lambda,$$

meaning that the volume of the problematic set of parameter values  $\beta \in [-1, 1]^p$  is given by

$$1 - (1 - 2\lambda)^{|S|}.$$

The cardinality  $|S|$  is the analogue of the number of edges in a DAG; for trees, the number of edges is  $p-1 \asymp p$  and hence, the comparable behavior for strong-faithfulness of trees and the volume of coefficients where the “beta-min” condition holds.

Using the lower bound computed in Theorem 5.2, we can also analyze some scaling of  $n$ ,  $p = p_n$  and  $\deg(G) = \deg(G_n)$  as a function of  $n$ , such that  $\lambda = \lambda_n$ -strong-faithfulness holds. This is discussed in Section 5.1.

We now provide a lower bound for DAGs where the skeleton is a cycle on  $p$  nodes.

**Theorem 5.3** (Lower bound for cycles). *Let  $C_p$  be a directed cycle on  $p$  nodes with edge set  $E$  as shown in Figure 3(b). Then*

$$\begin{aligned} (i) \quad & \frac{\text{vol}(\mathcal{M}_{C_p, \lambda})}{2^{|E|}} \geq 1 - (1 - \lambda)^{p + \binom{p-1}{2}}, \\ (ii) \quad & \frac{\text{vol}(\mathcal{N}_{C_p, \lambda}^{(1)})}{2^{|E|}} \geq 1 - (1 - \lambda)^{3p-2}, \\ (iii) \quad & \frac{\text{vol}(\mathcal{N}_{C_p, \lambda}^{(2)})}{2^{|E|}} \geq 1 - (1 - \lambda)^{2p-1}. \end{aligned}$$

For cycles, the measure of  $\lambda$ -strong-unfaithful distributions converges to 1 exponentially in  $p^2$ . The addition of a single cycle significantly increases the volume of strong-unfaithful distributions. The measure of restricted  $\lambda$ -strong-unfaithful distributions, however, converges to 1 exponentially in  $3p$  and hence shows a similar behavior as for trees. The scaling for achieving strong-faithfulness for cycles is discussed in Section 5.1.

*Proof.* Similar as for trees, all coordinate hyperplanes correspond to unfaithful distributions. The corresponding volume of strong-unfaithful distributions is  $2^{p-1} \cdot (2\lambda)$  and there are  $p$  such fattened hyperplanes. In addition, there are  $\binom{p-1}{2}$  hypersurfaces in the case of (i),  $2(p-1)$  hypersurfaces for (ii), and  $p-1$  hypersurfaces for (iii) defined by polynomials of the form  $f(\bar{a})a_{i,i+1} - g(\bar{a})a_{j,j+1}$ , where  $\bar{a} = \{a_{st} \mid (s, t) \notin \{(i, i+1), (j, j+1)\}\}$ . Such hypersurfaces are equivalently defined by

$$a_{i,i+1} = \frac{g(\bar{a})}{f(\bar{a})} a_{j,j+1}.$$

Since for any fixed  $\bar{a} \in [-1, 1]^{p-2}$  this is the parametrization of a line, we can lower bound the surface area of this hypersurface by  $2^{p-2} \cdot 2$ , which is the same lower bound as for a coordinate hyperplane. Similarly as in the proof for trees, an inclusion-exclusion argument over all hyperplanes yields the proof.  $\square$

Our simulations in Section 6 show that by increasing the number of cycles in the skeleton, the volume of strong-unfaithful distributions increases significantly. We now provide a lower bound for DAGs where the skeleton is a bipartite graph  $K_{2,p-2}$  and therefore consists of many 4-cycles. The corresponding scaling for strong-faithfulness is discussed in Section 5.1.

**Theorem 5.4** (Lower bound for bipartite graphs). *Let  $K_{2,p-2}$  be a directed bipartite graph on  $p$  nodes with edge set  $E$  as shown in Figure 3(c). Then*

$$(i) \quad \frac{\text{vol}(\mathcal{M}_{K_{2,p-2}, \lambda})}{2^{|E|}} \geq 1 - (1 - \lambda)^{(p-2)(2^{p-3}+1)},$$

$$\begin{aligned}
 (ii) \quad & \frac{\text{vol}(\mathcal{N}_{K_{2,p-2},\lambda}^{(1)})}{2^{|E|}} \geq 1 - (1 - \lambda)^{(p-2)(2^{p-3}+1)}. \\
 (iii) \quad & \frac{\text{vol}(\mathcal{N}_{K_{2,p-2},\lambda}^{(2)})}{2^{|E|}} \geq 1 - (1 - \lambda)^{(p-2)(2^{p-3}+1)}.
 \end{aligned}$$

*Proof.* The graph  $K_{2,p-2}$  has  $2(p-2)$  edges leading to  $2(p-2)$  hyperplanes of surface area  $2^{2(p-2)-1}$ . In addition, there are  $(p-2)(2^{p-3}-1)$  distinct hypersurfaces defined by polynomials of the form  $f(\bar{a})a_{1,j} - g(\bar{a})a_{j,p}$ . Their surface area can be lower bounded as well by  $2^{2(p-2)-1}$  as seen in the proof of Theorem 5.3. Hence, the volume of restricted and ordinary  $\lambda$ -strong-unfaithful distributions on  $K_{2,p-2}$  is bounded below by

$$1 - (1 - \lambda)^{2(p-2)+(p-2)(2^{p-3}-1)}.$$

□

## 5.1 Scaling and strong-faithfulness

We here consider the setting where the DAG  $G = G_n$  and hence the number of nodes  $p = p_n$  and the degree of the DAG  $\deg(G) = \deg(G_n)$  depend on  $n$ , and we take an asymptotic view point where  $n \rightarrow \infty$ . In such a setting, we focus on  $\lambda = \lambda_n \asymp \sqrt{\deg(G_n) \log(p_n)/n}$  (see [5]). We now briefly discuss when (restricted)  $\lambda_n$ -strong-faithfulness will asymptotically hold. For the latter, we must have that the lower bounds (see Theorems 5.2–5.4) on failure of (restricted)  $\lambda_n$ -strong-faithfulness tend to zero.

*Case I: lower bound*  $\asymp 1 - (1 - \lambda_n)^{p_n}$ . Such lower bounds appear for trees (Theorem 5.2) as well as for restricted strong-faithfulness for cycles (Theorem 5.3). The lower bound  $1 - (1 - \lambda_n)^{p_n}$  tends to zero as  $n \rightarrow \infty$  if

$$p_n = o\left(\sqrt{\frac{n}{\deg(G_n) \log(n)}}\right) \quad (n \rightarrow \infty).$$

Thus, we have  $p_n = o(\sqrt{n/\log(n)})$  for  $\lambda_n$ -strong-faithfulness for bounded degree trees and for restricted  $\lambda_n$ -strong faithfulness for cycles, and we have  $p_n = o((n/\log(n))^{1/3})$  for star-shaped graphs.

*Case II: lower bound*  $\asymp 1 - (1 - \lambda_n)^{p_n^2}$ . Such a lower bound appears for strong-faithfulness for cycles (Theorem 5.3). The lower bound  $1 - (1 - \lambda_n)^{p_n^2}$  tends to zero as  $n \rightarrow \infty$  if

$$p_n = o\left(\left(\frac{n}{\deg(G_n) \log(n)}\right)^{1/4}\right) \quad (n \rightarrow \infty).$$

Therefore, we have  $p_n = o((n/\log(n))^{1/4})$  for  $\lambda_n$ -strong-faithfulness for cycles.

*Case III: lower bound*  $\asymp 1 - (1 - \lambda_n)^{2^{p_n}}$ . This lower bound appears for strong-faithfulness for bipartite graphs (Theorem 5.4). This bound tends to zero as  $n \rightarrow \infty$  if

$$p_n = o(\log(n)) \quad (n \rightarrow \infty),$$

regardless of  $\deg(G_n) \leq p_n$ . Thus, for bipartite graphs with  $\deg(G_n) = p_n - 2$  we have  $p_n = o(\log(n))$  for  $\lambda_n$ -strong-faithfulness.



In summary, even for trees, we cannot have  $p_n \gg n$ , and high-dimensional consistency of the PC-algorithm seems rather unrealistic (unless e.g. the causal parameters have a distribution which is very different from uniform).

## 6 Simulation results

In this section, we describe various simulation results to validate the theoretical bounds described in the previous section. For our simulations we used the R library `pcalg` [6].

In a first set of simulations, we generated random DAGs with a given expected neighborhood size (i.e., expected degree of each vertex in the DAG) and edge weights sampled uniformly in  $[-1, 1]$ . We then analyzed how the proportion of  $\lambda$ -strong-unfaithful distributions depends on the number of nodes  $p$  and the expected neighborhood size of the graph. Depending on the number of nodes in a graph, we analyzed 5-10 different expected neighborhood sizes and generated 10,000 random DAGs for each expected neighborhood size.

Using `pcalg` we computed all partial correlations. Since this computation requires multiple matrix inversions, numerical imprecision has to be expected. We assumed that all partial correlations smaller than  $10^{-12}$  were actual zeroes and counted the number of simulations, for which the minimal partial correlation (after excluding the ones with partial correlation  $< 10^{-12}$ ) was smaller than  $\lambda$ . The resulting plots of the proportion of  $\lambda$ -strong-unfaithful distributions for three different values of  $\lambda$ , namely  $\lambda = 0.1, 0.01, 0.001$  are given in Figure 4(a) for  $p = 3$  nodes, in Figure 4(b) for  $p = 5$  nodes and in Figure 4(c) for  $p = 10$  nodes.

It appears that already for very sparse graphs (i.e., expected neighborhood size of 2) and relatively small graphs (i.e., 10 nodes) the proportion of  $\lambda$ -strong-unfaithful distributions is nearly 1 for  $\lambda = 0.1$ , about 0.9 for  $\lambda = 0.01$  and about 0.7 for  $\lambda = 0.001$ . In addition, the proportion of  $\lambda$ -strong-unfaithful distributions increases with graph density and with the number of nodes (even for a fixed expected neighborhood size). The general upper bound derived in Theorem 5.1 shows similar behaviors. The number of summands and the degrees of the hypersurfaces grow with the number of nodes and graph density.

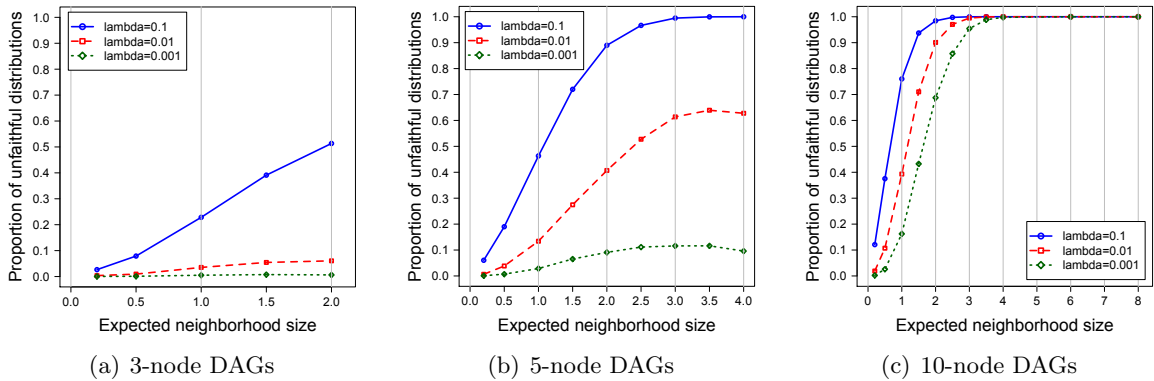


Figure 4: Proportion of  $\lambda$ -strong-unfaithful distributions for 3 values of  $\lambda$ .

# GEOMETRY OF FAITHFULNESS ASSUMPTION IN CAUSAL INFERENCE

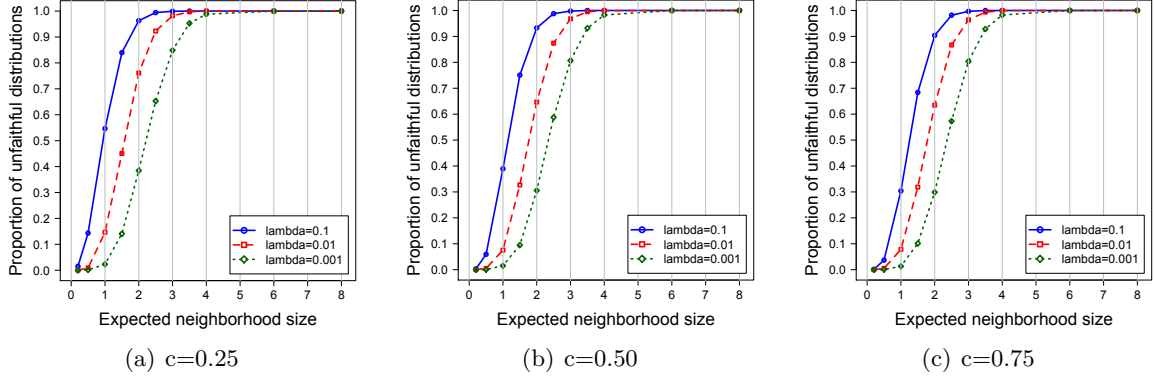


Figure 5: Proportion of  $\lambda$ -strong-unfaithful distributions for 10-node DAGs when restricting the parameter space.

## 6.1 Bounding away the causal parameters from zero

In the following, we analyze how the proportion of  $\lambda$ -strong-unfaithful distributions changes when restricting the parameter space. The motivation behind this experiment is that unfaithfulness would not be too serious of an issue if the PC-algorithm only fails to recover very small causal effects but does well when the causal parameters are large. We repeated the experiments when restricting the parameter space to

$$[-1, -c] \cup [c, 1]$$

for  $c = 0.25, 0.5$  and  $0.75$ . The results for 10-node DAGs are shown in Figure 5. Restricting the parameter space seems to help for sparse graphs but doesn't seem to play a role for dense graphs. We now analyze various classes of graphs and their behavior when restricting the parameter space.

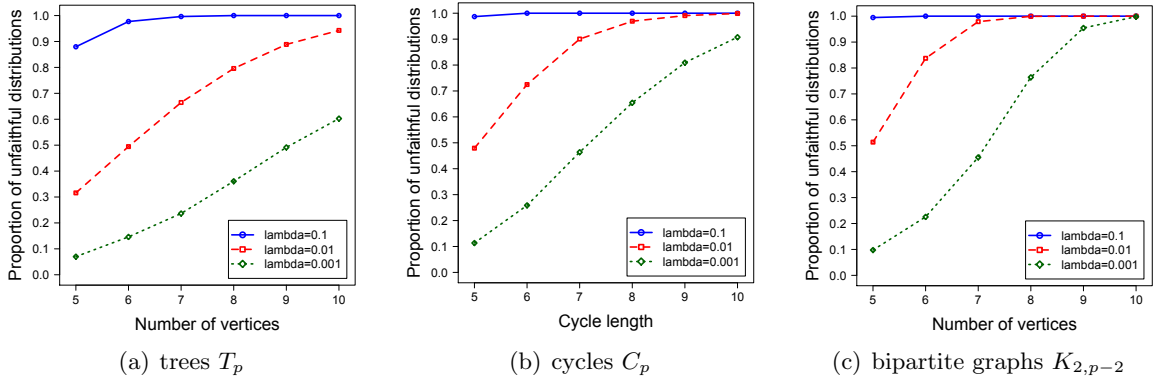


Figure 6: Proportion of  $\lambda$ -strong-unfaithful distributions when the skeleton is a tree, a cycle or a bipartite graph.

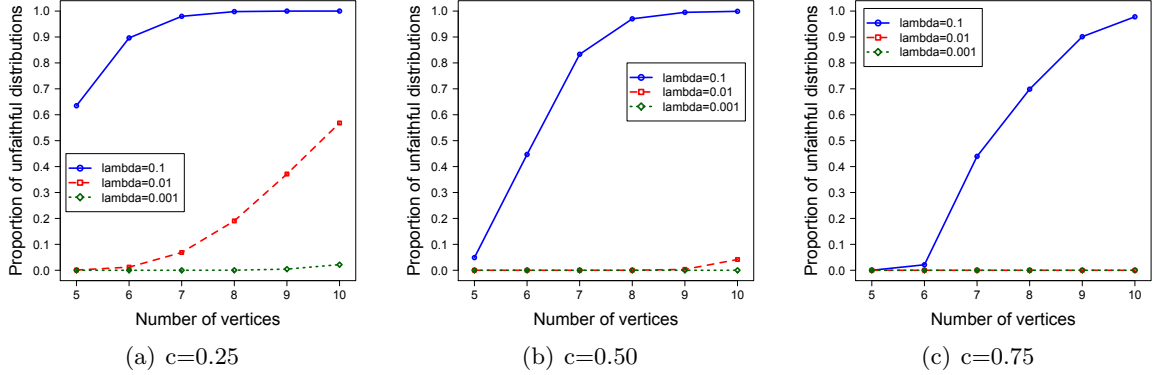


Figure 7: Proportion of  $\lambda$ -strong-unfaithful distributions for trees when restricting the parameter space.

### 6.1.1 Trees

We generated connected trees where all edges are directed away from the root by first sampling the number of levels uniformly from  $\{2, \dots, p\}$  (a tree with 2 levels is a star graph, a tree with  $p$  levels is a line), then distributing the  $p$  nodes on these levels such that there is at least one node on each level, and finally assigning a unique parent to each node uniformly from all nodes on the previous level. The resulting plots for the whole parameter space  $[-1, 1]$  are shown in Figure 6(a). The plots when restricting the parameter space for  $c = 0.25, 0.5$  and  $0.75$  are shown in Figure 7. As before, each proportion is computed from 10,000 simulations.

For trees restricting the parameter space reduces the proportion of  $\lambda$ -strong-unfaithful distributions by a large amount. This can be explained by the special structure of the defining polynomials (given in Corollary 4.3). Since the defining polynomials of the partial correlation hypersurfaces are of the form  $a_{i \rightarrow j} \cdot (1 + \text{SOS}(a))$ , the minimal possible value of these polynomials when restricting the parameter space is

$$c^{\text{path length from } i \text{ to } j}.$$

### 6.1.2 Cycles

We generated DAGs where the skeleton is a cycle and the edges are directed as shown in Figure 3(b). The edge weights were sampled uniformly from  $[-1, -c] \cup [c, 1]$ . The resulting plots for the whole parameter space are shown in Figure 6(b). The plots for the restricted parameter space with  $c = 0.25, 0.5$  and  $0.75$  are shown in Figure 8. Again, each point corresponds to 10,000 DAGs.

For cycles restricting the parameter space also reduces the proportion of  $\lambda$ -strong-unfaithful distributions, however not as drastically as for trees. This can again be explained by the special structure of the defining polynomials (given in Corollary 4.3). When the defining polynomials are of the form  $f(\bar{a})a_{i,i+1} - g(\bar{a})a_{j,j+1}$ , they might evaluate to a very small number even when the parameters themselves are large.

## GEOMETRY OF FAITHFULNESS ASSUMPTION IN CAUSAL INFERENCE

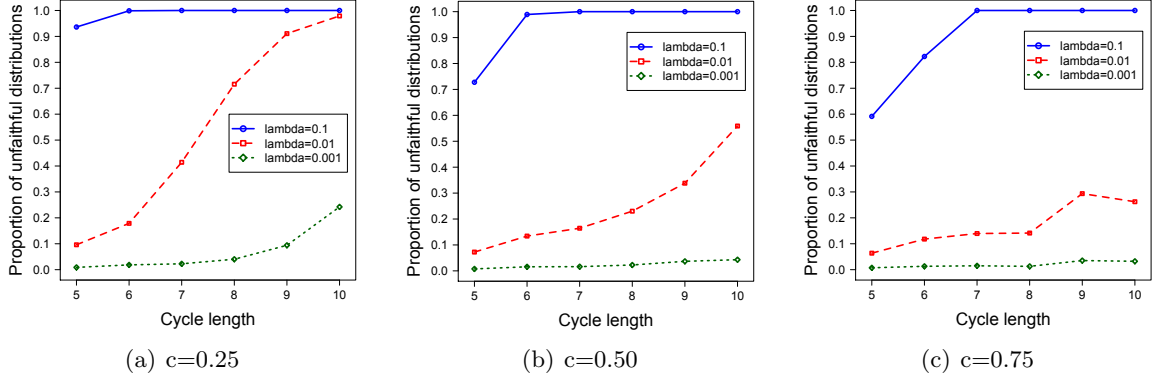


Figure 8: Proportion of  $\lambda$ -strong-unfaithful distributions for cycles when restricting the parameter space.

### 6.1.3 Bipartite graphs

We generated DAGs where the skeleton is a bipartite graph  $K_{2,p-2}$  and the edges are directed as shown in Figure 3(c). Bipartite graphs  $K_{2,p-2}$  consist of many 4-cycles. For such graphs there are many paths from one vertex to another and therefore many ways for a polynomial to cancel out, even when the parameter values are large. As a consequence, for such graphs restricting the parameter space makes hardly no difference on the proportion of  $\lambda$ -strong-unfaithful distributions. This becomes apparent in Figure 6(c) and Figure 9.

### 6.1.4 Lower bounds

We compare the theoretical lower bounds derived in Section 5 to the simulation results in this section for DAGs where the skeleton is a tree, a cycle or a bipartite graph when  $c = 0$ . We present our lower bounds together with the simulation results in Figure 10. The black lines

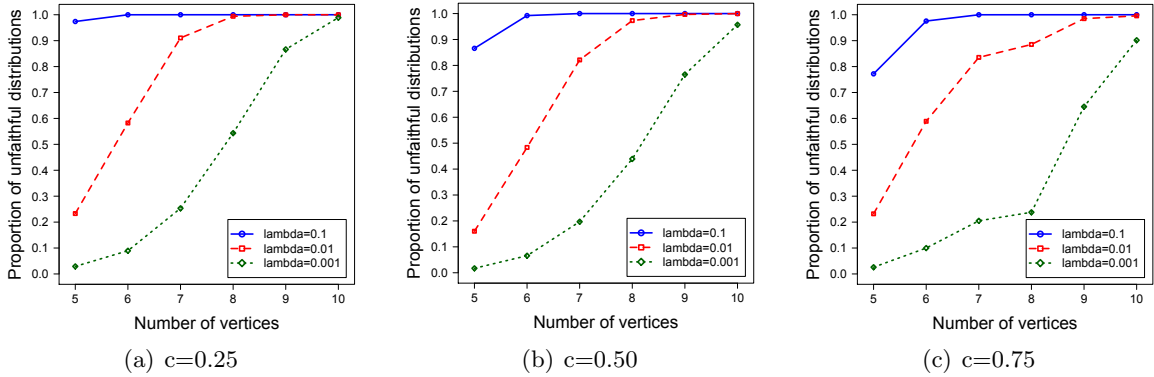


Figure 9: Proportion of  $\lambda$ -strong-unfaithful distributions for bipartite graphs  $K_{2,p-2}$  when restricting the parameter space.

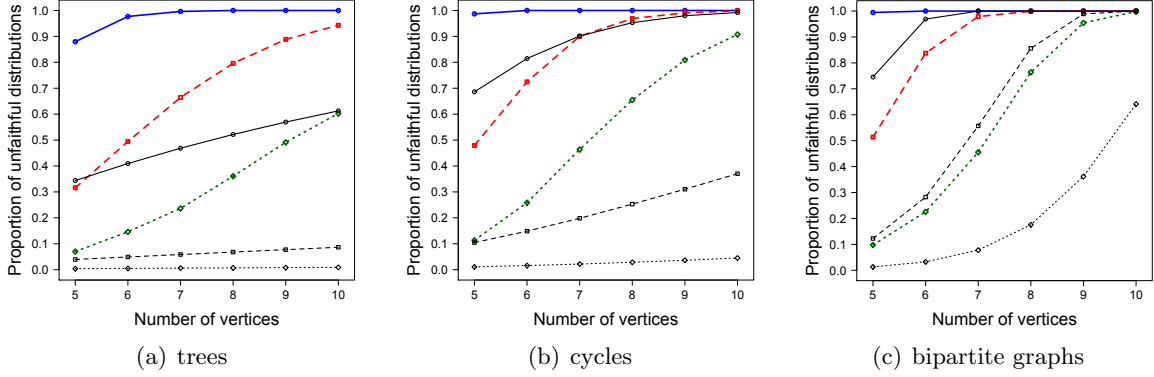


Figure 10: Comparison of theoretical lower bounds and approximated proportion of  $\lambda$ -strong-unfaithful distributions for trees, cycles and bipartite graphs  $K_{2,p-2}$ .

correspond to the lower bounds, the solid line to  $\lambda = 0.1$ , the dashed line to  $\lambda = 0.01$  and the dotted line to  $\lambda = 0.001$ . In particular for bipartite graphs our lower bounds approximate the simulation results very well.

## 6.2 Restricted $\lambda$ -strong-faithfulness

As already discussed earlier, the PC-algorithm only requires the computation of all partial correlations over edges in the graph  $G$  and conditioning sets  $S$  of size at most  $\deg(G)$ . In order to analyze when the (conservative) PC-algorithm works, we repeated all our simulations when restricting the partial correlations to edges in the graph  $G$  and conditioning sets  $S$  of size at most  $\deg(G)$ , i.e., part (i) of the restricted strong-faithfulness assumption in Definition 1.4, called the adjacency-faithfulness assumption. The results for general 10-node DAGs are shown in Figure 11. We see that the proportion of  $\lambda$ -adjacency-unfaithful distributions is slightly reduced compared to the proportion of  $\lambda$ -strong-unfaithful distributions shown in Figure 5, in particular for sparse graphs. For trees and bipartite graphs the

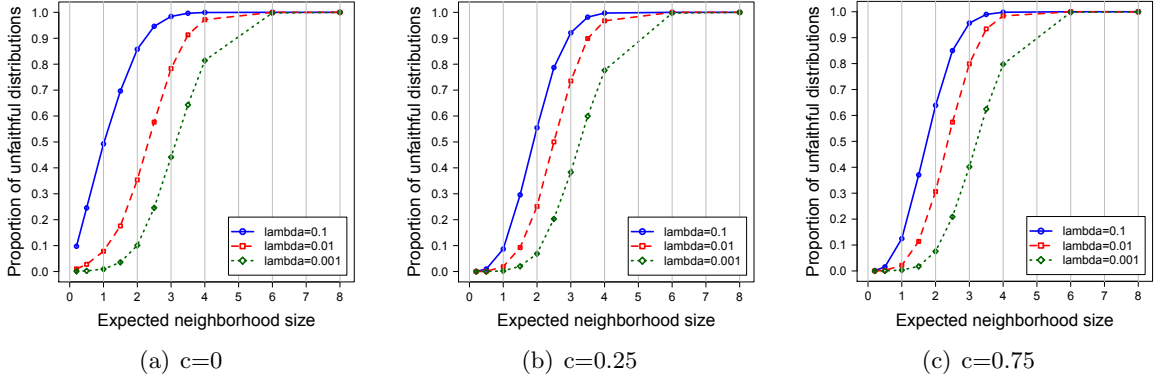


Figure 11: Proportion of  $\lambda$ -adjacency-unfaithful distributions for 10-node DAGs.

proportion of restricted  $\lambda$ -strong-unfaithful distributions is similar to the proportion of  $\lambda$ -strong-unfaithful distributions shown in Figures 6, 7 and 9, whereas the behavior for cycles regarding the proportion of restricted  $\lambda$ -strong-unfaithful distributions is similar to trees. We don't repeat these plots here, but we remark that they nicely agree with the theoretical bounds for restricted  $\lambda$ -strong-faithfulness and  $\lambda$ -adjacency-faithfulness derived in Section 5.

## 7 Discussion

In this paper, we have shown that the (restricted) strong-faithfulness assumption is very restrictive, even for relatively small and sparse graphs. Furthermore, the proportion of strong-unfaithful distributions grows with the number of nodes and the number of edges. We have also analyzed the restricted strong-faithfulness assumption introduced by Spirtes and Zhang [15], a weaker condition than strong-faithfulness, which is essentially a necessary condition for uniform or high-dimensional consistency of the popular PC-algorithm and of the conservative PC-algorithm. As seen in this paper, our lower bounds on restricted strong-unfaithful distributions are similar to our bounds for strong faithfulness, implying inconsistent estimation with the PC-algorithm for a relatively large class of DAGs.

For trees, due to the special structure of the polynomials defining the hypersurfaces of unfaithful distributions, if the causal parameters are large, the partial correlations tend to stay away from these hypersurfaces and strong-faithfulness holds for a large proportion of distributions. However, as soon as there are cycles in the graph (even for sparse graphs), the polynomials can cancel out also for large causal parameters, and the strong-faithfulness assumption does not hold. More precisely, if the skeleton is a single cycle, our lower bounds on the proportion of restricted strong-unfaithful distributions is of the same order of magnitude as for trees. However, if the skeleton consists of multiple cycles as for example for bipartite graphs, the lower bounds for restricted strong-unfaithful distributions are as bad as for plain strong-unfaithful distributions.

Assuming our framework and in view of the discussion above, in the presence of cycles in the skeleton, the (conservative) PC-algorithm is not able to consistently estimate the true underlying Markov equivalence class when  $p$  is large relative to  $n$ , even for large causal parameters (large edge weights). Some special assumptions on the sparsity and causal parameters might help, but without making such assumptions, the limitation is in the range where  $p = p_n = o(\sqrt{n/\log(n)})$ . This constitutes a severe limitation of the PC-algorithm. As an alternative method, the penalized maximum likelihood estimator [3, cf.] does not require strong-faithfulness but instead a stronger version of a beta-min condition (i.e., sufficiently large causal parameters) [13] which seems weaker than strong-faithfulness. In view of this, our presented results on strong-faithfulness indicate an advantage of the penalized maximum likelihood estimator over the PC-algorithm.

Throughout the paper we have assumed that the causal parameters are uniformly distributed in the hypercube  $[-1, 1]^{|E|}$ . Since all hypersurfaces corresponding to unfaithful distributions go through the origin, a prior distribution which puts more mass around the origin (e.g. a Gaussian distribution) would lead to a higher proportion of strong-unfaithful distributions, whereas a prior distribution which puts more mass on the boundary of the hy-

percube  $[-1, 1]$  would reduce the proportion of strong-unfaithful distributions. Computing and comparing these measures for different priors would be an interesting extension of our work.

## 8 Proofs

*Proof of Proposition 4.1.* Statement (i) follows from the matrix inversion formula using the cofactor matrix, i.e.,

$$\Sigma_{ij} = \frac{1}{\det(K)} C(K)_{ij},$$

and the fact that the concentration matrix  $K$  is positive definite and therefore  $\det(K) > 0$ . Statement (ii) is a well-known fact about the multivariate Gaussian distribution.

Let  $A, B \subset V$  be two subsets of vertices. We denote by  $K_{AB}$  the submatrix of  $K$  consisting of the entries  $K_{ij}$ , where  $(i, j) \in A \times B$ . Let  $K_A$  denote the concentration matrix in the Gaussian model, where we marginalized over  $A^c = V \setminus A$ . With these definitions we have that

$$K_A = \Sigma_{AA}^{-1}.$$

The correlation between  $X_i$  and  $X_j$  conditioned on  $S$  corresponds to the  $(i, j)$ -th entry in the matrix  $K_Q$ . Using the Schur complement formula, we get that

$$K_Q = K_{QQ} - K_{QQ^c} (K_{Q^c Q^c})^{-1} K_{Q^c Q}. \quad (12)$$

Since  $K_{Q^c Q^c}$  is positive definite, we can rewrite Equation (12) as

$$\det(K_{Q^c Q^c}) K_Q = \det(K_{Q^c Q^c}) K_{QQ} - K_{QQ^c} C(K_{Q^c Q^c}) K_{Q^c Q},$$

from which statement (iii) follows.  $\square$

*Proof of (10).* We first note that the  $(i, j)$ -th element of  $A^s$  consists of the sum of the weights of all paths  $p = (p_0, p_1, \dots, p_s)$  with  $p_0 = i$  and  $p_s = j$  for which  $(p_{k-1}, p_k) \in E$  for all  $k = 1, \dots, s$ . This means that  $(A^s)_{ij}$  corresponds to all "forward" paths from  $i$  to  $j$  of length  $s$ . Analogously,  $(A^T)^r$  corresponds to all "backward" paths from  $i$  to  $j$  of length  $r$ .

We decompose the covariance matrix using the Neumann power series. We can do this since all eigenvalues of the matrix  $A$  are zero (because  $A$  is upper triangular).

$$\begin{aligned} \Sigma &= ((I - A)(I - A)^T)^{-1} \\ &= \sum_{k=0}^{\infty} \sum_{r+s=k} (A^T)^r A^s \\ &= \sum_{k=0}^{2p-2} \sum_{\substack{r+s=k, \\ r, s \leq p-1}} (A^T)^r A^s. \end{aligned}$$

For the last inequality we used the assumption that the underlying graph is acyclic. Using the path interpretation it is clear that for acyclic graphs the matrix  $A^s$  is the zero-matrix for all  $s \geq p$ .  $\square$

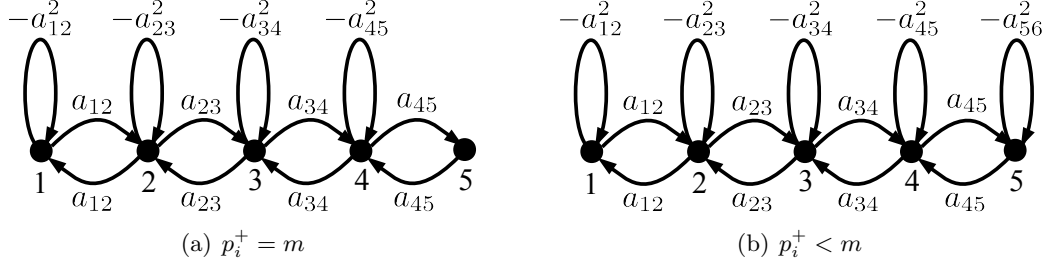


Figure 12: Subgraphs  $\hat{G}_{P_i}$ , where  $G$  is a directed line and  $P_i = \{1, 2, \dots, 5\}$ .

*Proof of Corollary 4.3.* To prove (a) we first consider the special case where  $G$  is a directed line on  $p$  nodes, where all edges point in the same direction, i.e.,  $(i, i+1) \in E$  for  $1 \leq i < p$ . The following argument can then easily be generalized to directed trees  $T_p$ .

Let  $i, j \in V$  and without loss of generality we assume that  $i < j$ . Since there are no colliders in  $G$ , it follows from (9) that

$$K_{ij} = \begin{cases} -a_{ij} & \text{if } j \text{ is a child of } i \\ 0 & \text{otherwise} \end{cases}$$

$\Sigma_{ij}$  corresponds to all collider-free paths from  $i$  to  $j$  and therefore

$$\Sigma_{ij} = (1 + a_{i-1,i}^2 (1 + a_{i-2,i-1}^2 (\dots (1 + a_{12}^2)))) \prod_{k=i}^{j-1} a_{k,k+1}. \quad (13)$$

The first term corresponds to the value of all collider-free loops from  $i$  to  $i$  and the second term to the value of the path from  $i$  to  $j$ .

Let  $S \subseteq V \setminus \{i, j\}$  and  $Q = S \cup \{i, j\}$ . If there exists an element  $s \in S$  such that  $i < s < j$ , then the CI relation  $X_i \perp\!\!\!\perp X_j \mid X_S$  is already entailed by the Markov condition. We can therefore assume without loss of generality that there is no  $s \in S$  such that  $i < s < j$ . Since there are no colliders in  $G$ , it follows from Proposition 4.1 (iii) that the corresponding polynomial is of the form

$$\begin{cases} -\det(K_{Q^c Q^c}) a_{ij} & \text{if } j \text{ is a child of } i \\ -\sum_{p,q \in Q^c} a_{ip} C(K_{Q^c Q^c})_{pq} a_{qj} & \text{otherwise} \end{cases} \quad (14)$$

The corresponding symmetrized and reweighted graph  $\hat{G}$  for  $p = 5$  is shown in Figure 12(a). Note that there is a unique self-avoiding path between any two vertices. As a consequence, the polynomial corresponding to the CI relation  $X_i \perp\!\!\!\perp X_j \mid X_S$  in (14) can be written as

$$-\left(1 + \sum_{k=1}^{|P|} \sum_{m_1 + \dots + m_s = k} (-1)^s \mu(c_{m_1}) \dots \mu(c_{m_s})\right) \prod_{k=i}^{j-1} a_{k,k+1}, \quad (15)$$

where  $P = Q^c \setminus \{i+1, \dots, j-1\}$ .



We now analyze the cycles in  $P$ . We decompose  $P$  into intervals  $P = P_1 \cup \dots \cup P_s$ , where  $P_i = \{p_i^-, p_i^- + 1, \dots, p_i^+\}$ . We need to distinguish two cases. If  $p_i^+ = p$ , then the subgraph  $\hat{G}_{P_i}$  is of the form as shown in Figure 12(a) (for  $p_i^- = 1$  and  $p_i^+ = 5$ ). Otherwise the subgraph is of the form as shown in Figure 12(b) (for  $p_i^- = 1$  and  $p_i^+ = 5$ ).

We note that all cycles are either of length 1 (with value  $-a_{k,k+1}^2$ ) or of length 2 (with value  $a_{k,k+1}^2$ ). In the case where  $p_i^+ = p$  all cycles of length 1 cancel with the cycles of length 2. In the case where  $p_i^+ < p$ , however, the cycle of length 1 with value  $-a_{p_i^+, p_i^++1}^2$  does not cancel and therefore neither does the combination of  $k$  cycles

$$\prod_{j=0}^{k-1} (-a_{p_i^+-j, p_i^+-j+1}^2)$$

for any  $k \in \{1, \dots, p_i^+ - p_i^-\}$ . As a consequence, the polynomial corresponding to the CI relation  $X_i \perp\!\!\!\perp X_j \mid X_S$  in (15) can be written as

$$-\prod_{i=1}^s \left( 1 + a_{p_i^+-1, p_i^+}^2 \left( 1 + a_{p_i^+-2, p_i^+-1}^2 \left( \dots \left( 1 + a_{p_i^-, p_i^-+1}^2 \right) \right) \right) \right) \prod_{k=i}^{j-1} a_{k,k+1}.$$

The proofs for (b) and (c) are analogous and basically require understanding the cycles in  $\hat{G}$ .  $\square$

## Acknowledgments

We wish to thank Marloes Maathuis and Mohab Safey El Din for helpful discussions. This work was supported in part by US NSF grants DMS-0907632, DMS-1107000, SES-0835531 (CDI) and ARO grant W911NF-11-1-0114. This work was also supported in part by the Center for Science of Information (CSoI), a US NSF Science and Technology Center, under grant agreement CCF-0939370.

## References

- [1] D. Geiger A. Becker and C. Meek. Perfect tree-like markovian distributions. *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 19–23, 2000.
- [2] R. J. Adler and J. E. Taylor. *Random fields and geometry*. Springer Monographs in Mathematics. Springer, New York, 2007.
- [3] D.M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.
- [4] L. Guth. Minimax problems related to cup powers and steenrod squares. *Geometric And Functional Analysis*, 18:1917–1987, 2008.

- [5] M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.
- [6] M. Kalisch, M. Mächler, D. Colombo, M.H. Maathuis, and P. Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* 47, pages 1–26, 2011.
- [7] M.H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37:3133–3164, 2009.
- [8] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.
- [9] J. Ponsstein. Self-avoiding paths and the adjacency matrix of graph. *SIAM Journal on Applied Mathematics*, 14:600–609, 1966.
- [10] J. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 90:491–515, 2003.
- [11] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, second edition, 2001.
- [12] S. Sullivant, K. Talaska, and J. Draisma. Trek separation for gaussian graphical models. *The Annals of Statistics*, 38:1665–1685, 2010.
- [13] S. van de Geer and P. Bühlmann. Penalized maximum likelihood estimation for sparse directed acyclic graphs, 2012. Preprint arXiv:1205.5473v1.
- [14] J. Zhang and P. Spirtes. Strong faithfulness and uniform consistency in causal inference. In *Uncertainty in Artificial Intelligence (UAI)*, pages 632–639, 2003.
- [15] J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271, 2008.